

Epigenetics and DNase-Seq

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2018

Anthony Gitter

gitter@biostat.wisc.edu

Goals for lecture

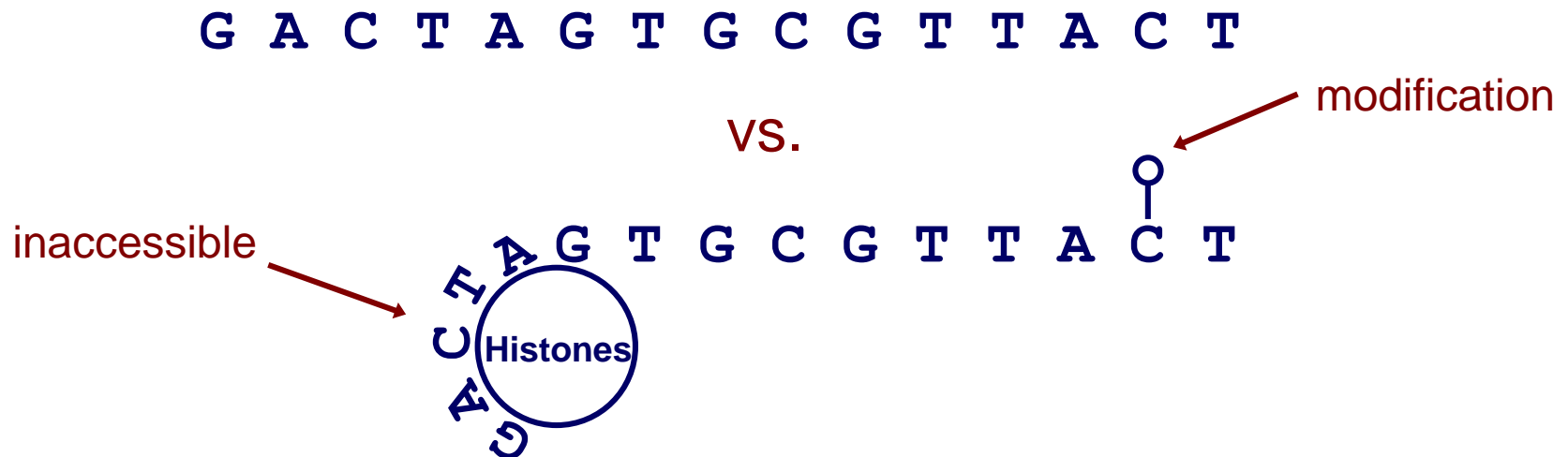
Key concepts

- Importance of epigenetic data for understanding transcriptional regulation
- Predicting transcription factor binding sites
- Gaussian process models

Introduction to epigenetics

Defining epigenetics

- Formally: attributes that are “in addition to” genetic sequence or sequence modifications
- Informally: experiments that reveal the context of DNA sequence
 - DNA has multiple states and modifications



Importance of epigenetics

Better understand

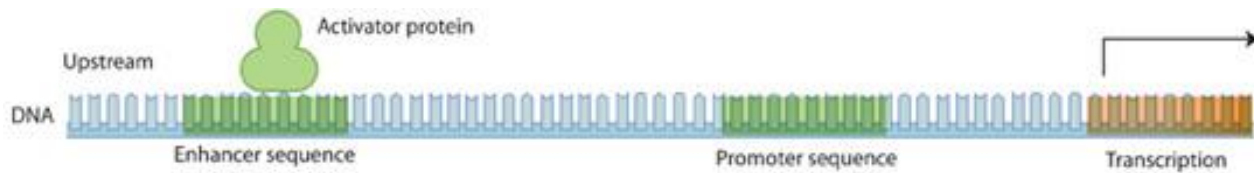
- DNA binding and transcriptional regulation
- Differences between cell and tissue types
- Development and other important processes
- Non-coding genetic variants (next lecture)

PWMs are not enough

- Genome-wide motif scanning is imprecise
- Transcription factors (TFs) bind < 5% of their motif matches
- Same motif matches in all cells and conditions

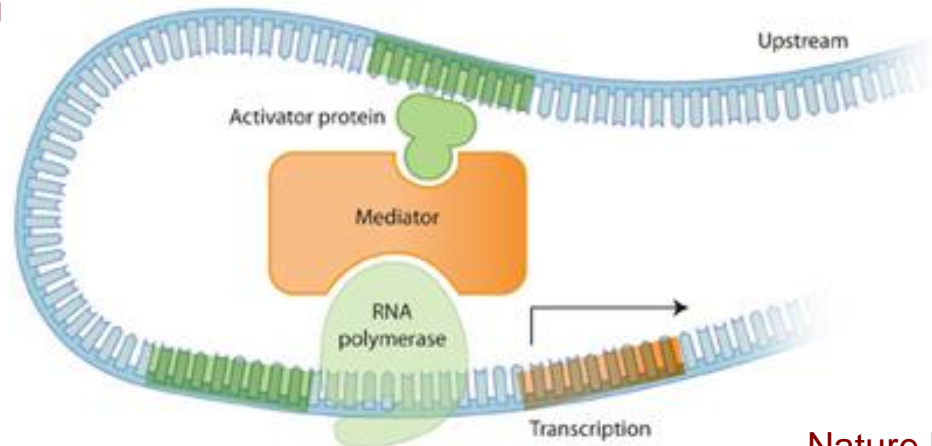
PWMs are not enough

- DNA looping can bring distant binding sites close to transcription start sites
- Which genes does an enhancer regulate?



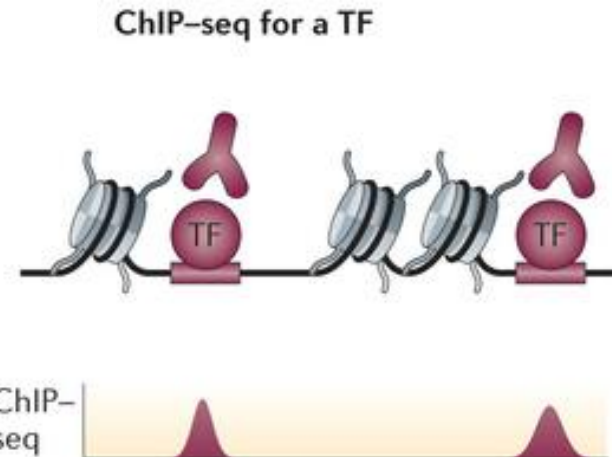
Enhancer: DNA binding site for TFs, can be far from affected gene

Promoter: DNA binding site for TFs, close to gene transcription start site



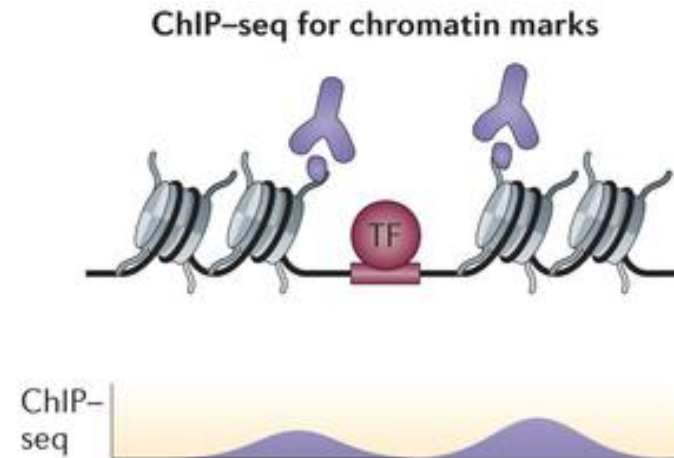
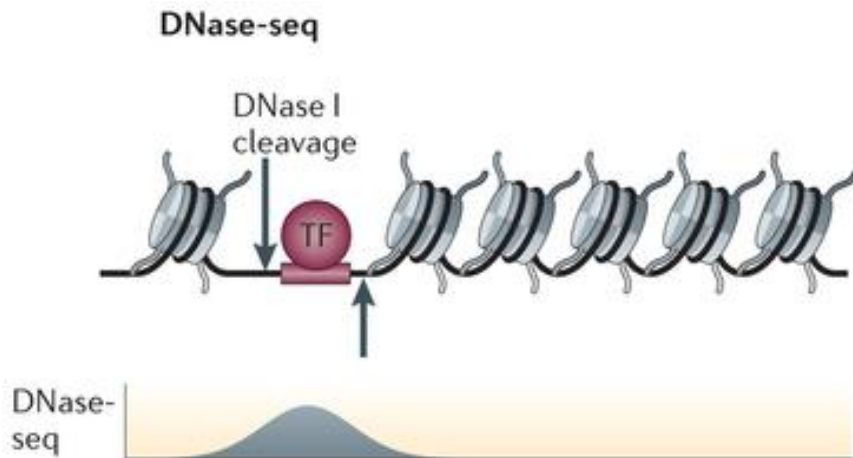
Mapping regulatory elements genome-wide

- Can do much better than motif scanning with additional data
- ChIP-seq measures binding sites for one TF at a time



Shlyueva *Nature Reviews Genetics* 2014

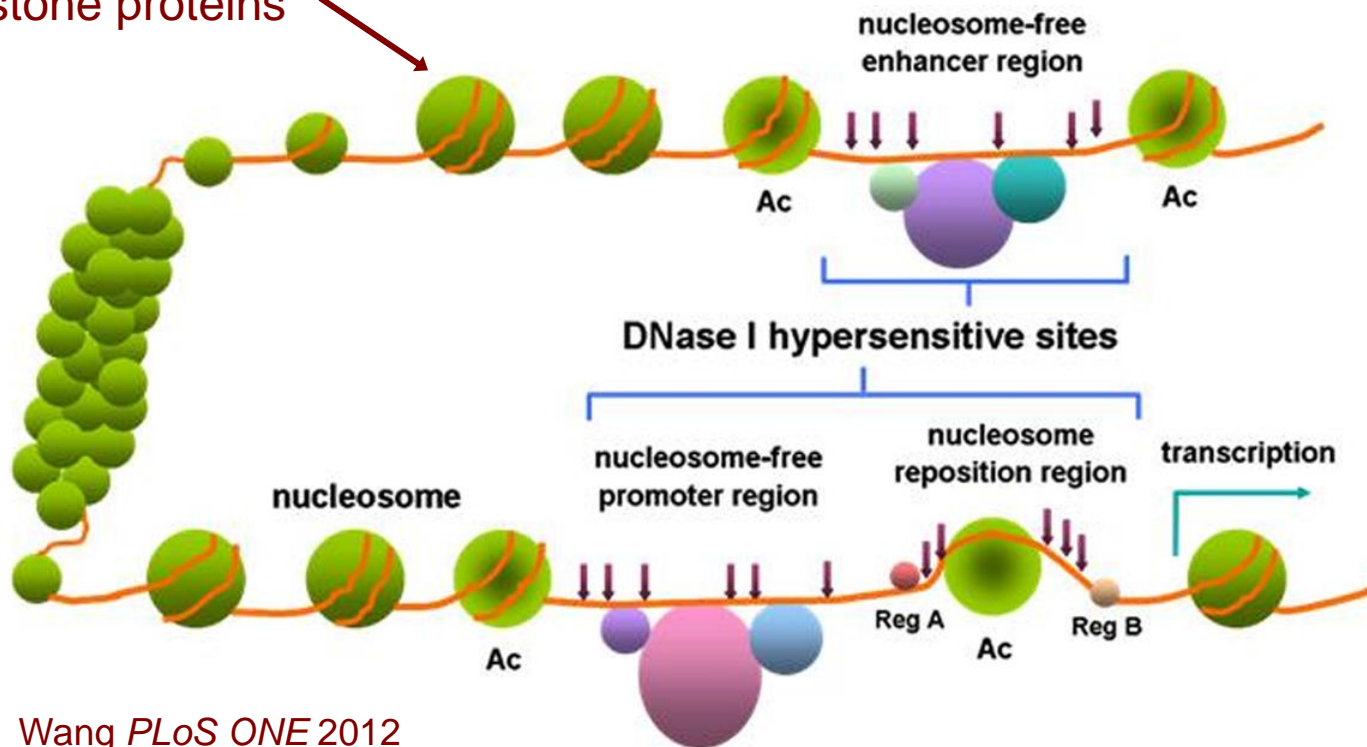
- Epigenetic data suggests where *some* TF binds



DNase I hypersensitivity

- Regulatory proteins bind accessible DNA
- DNase I enzyme cuts open chromatin regions that are not protected by nucleosomes

Nucleosome: DNA wrapped around histone proteins

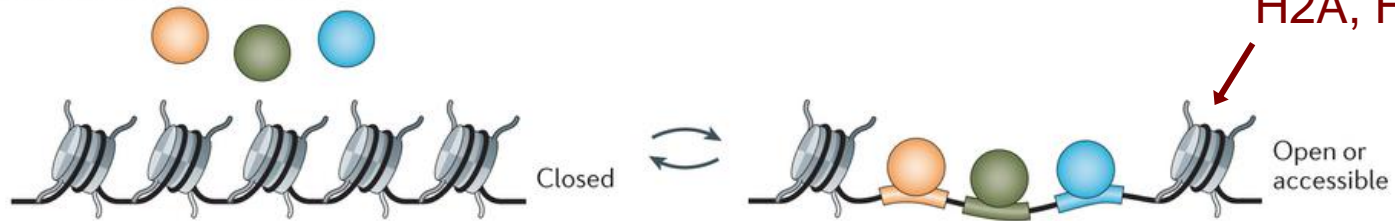


Histone modifications

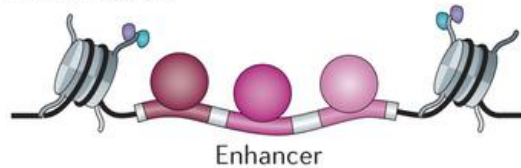
- Mark particular regulatory configurations

Two copies of histone proteins H2A, H2B, H3, H4

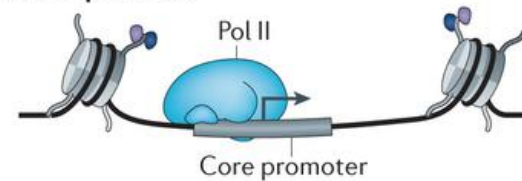
Chromatin as accessibility barrier



Active enhancer



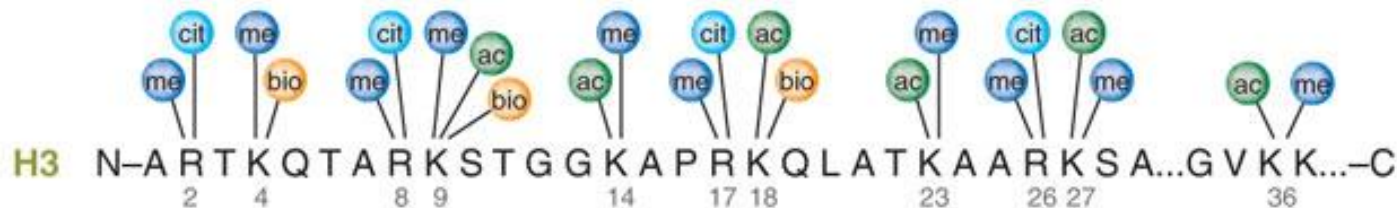
Active promoter



Shlyueva *Nature Reviews Genetics* 2014



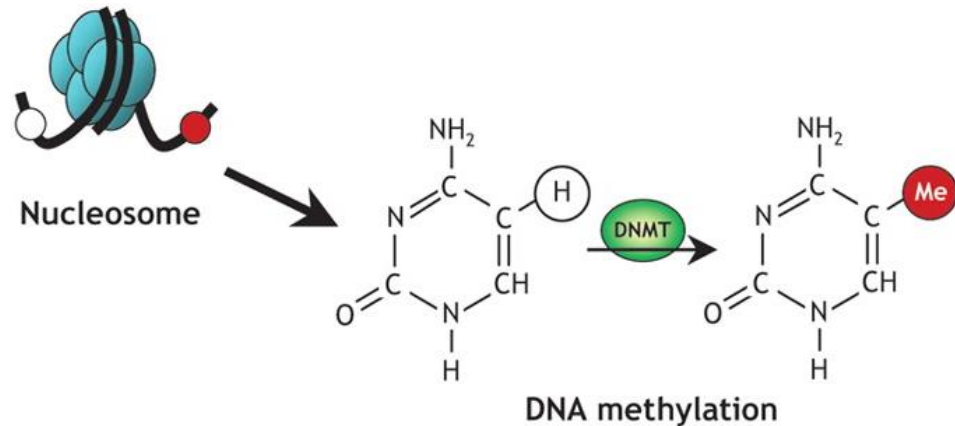
- H3 (protein) K27 (amino acid) ac (modification)



Latham *Nature Structural & Molecular Biology* 2007; Katie Ris-Vicari

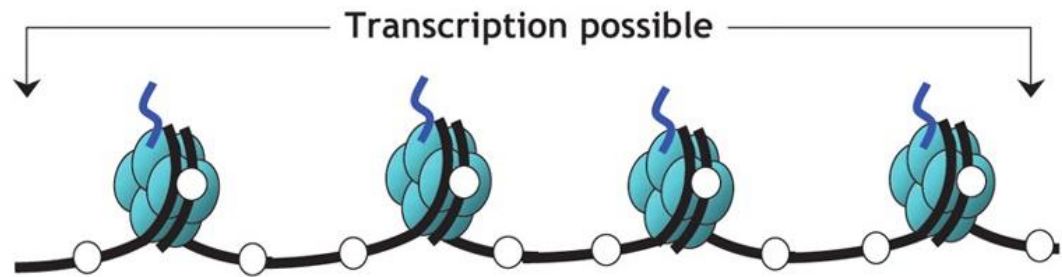
DNA methylation

- Reversible DNA modification
- Represses gene expression



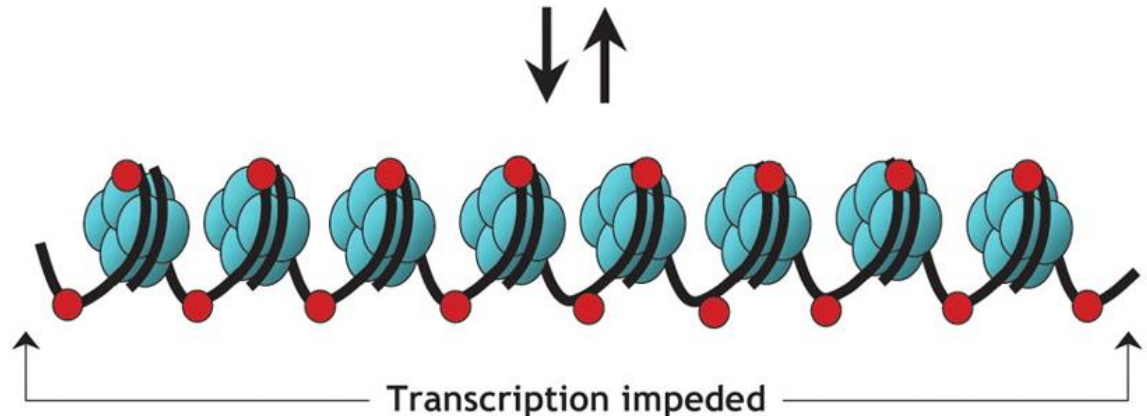
Gene “switched on”

- Active (open) chromatin
- Unmethylated cytosines (white circles)
- Acetylated histones



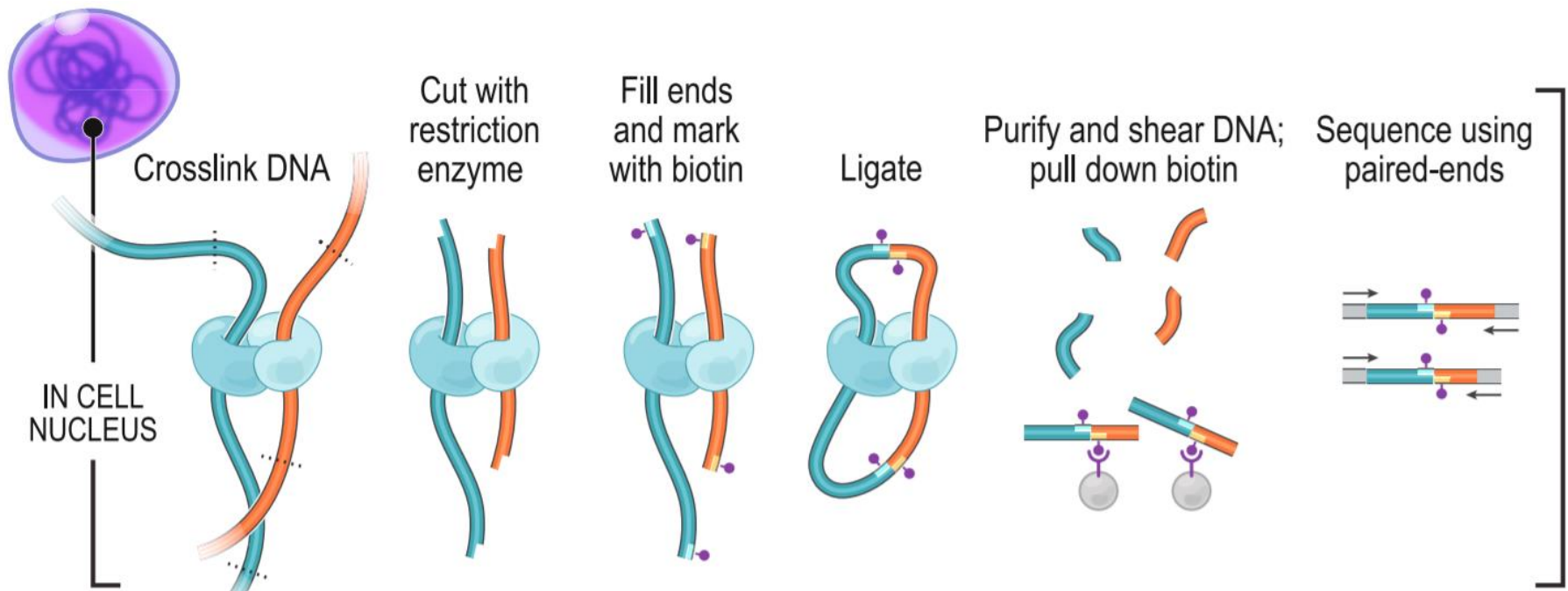
Gene “switched off”

- Silent (condensed) chromatin
- Methylated cytosines (red circles)
- Deacetylated histones

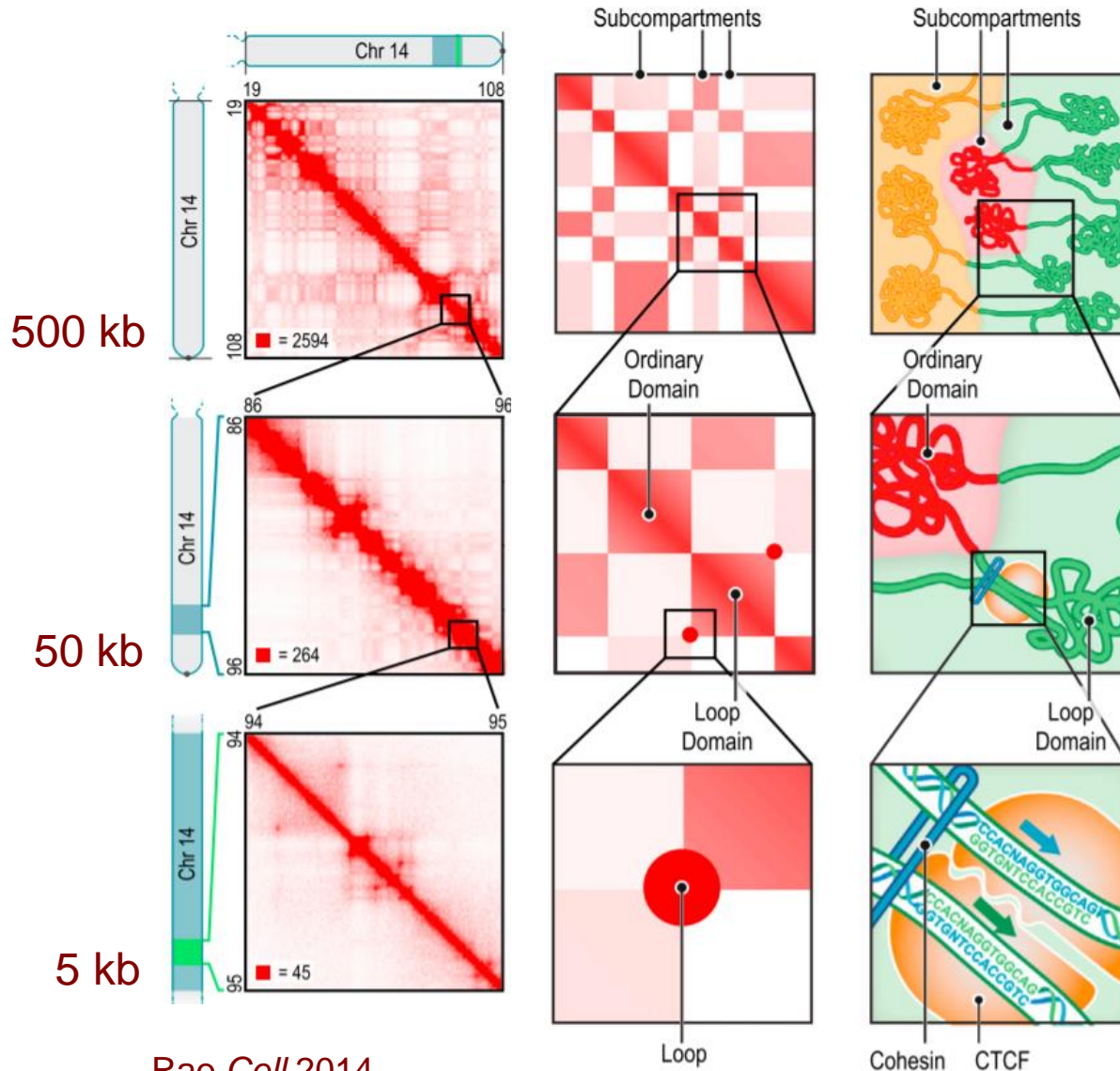


3d organization of chromatin

- Algorithms to predict long range enhancer-promoter interactions
- Or measure with chromosome conformation capture (3C, Hi-C, etc.)



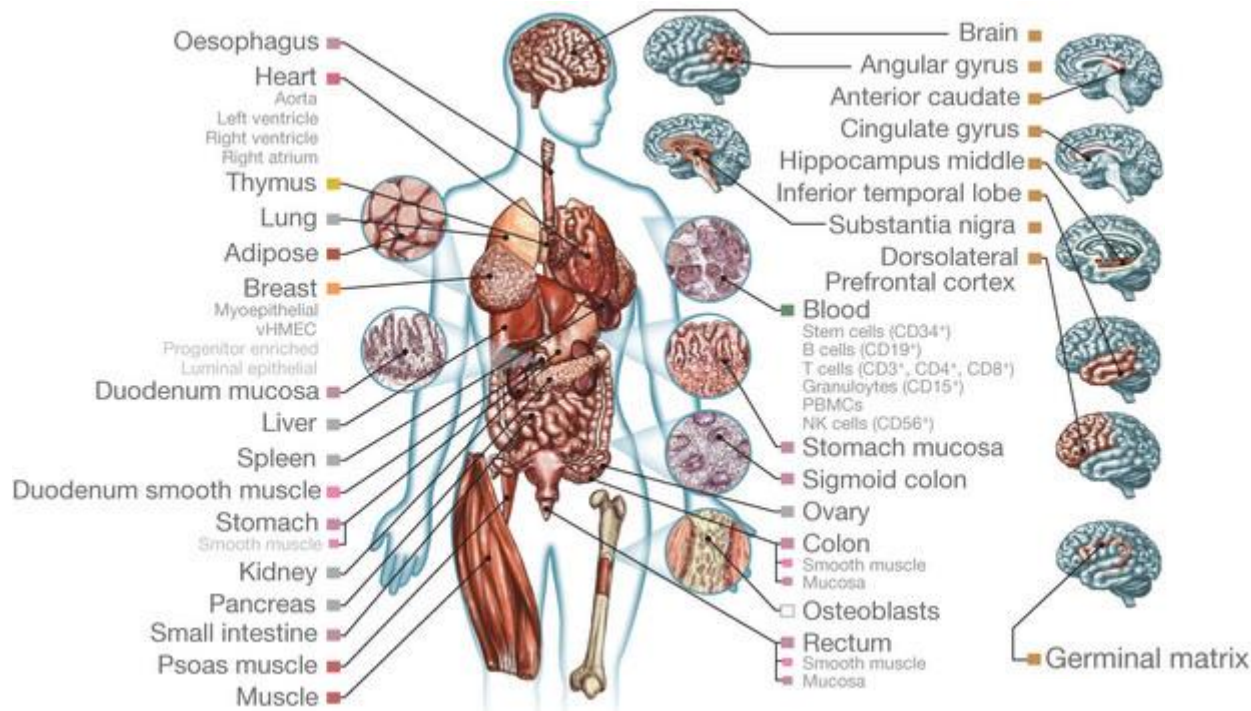
3d organization of chromatin



- Hi-C produces 2d chromatin contact maps
- Learn domains, enhancer-promoter interactions

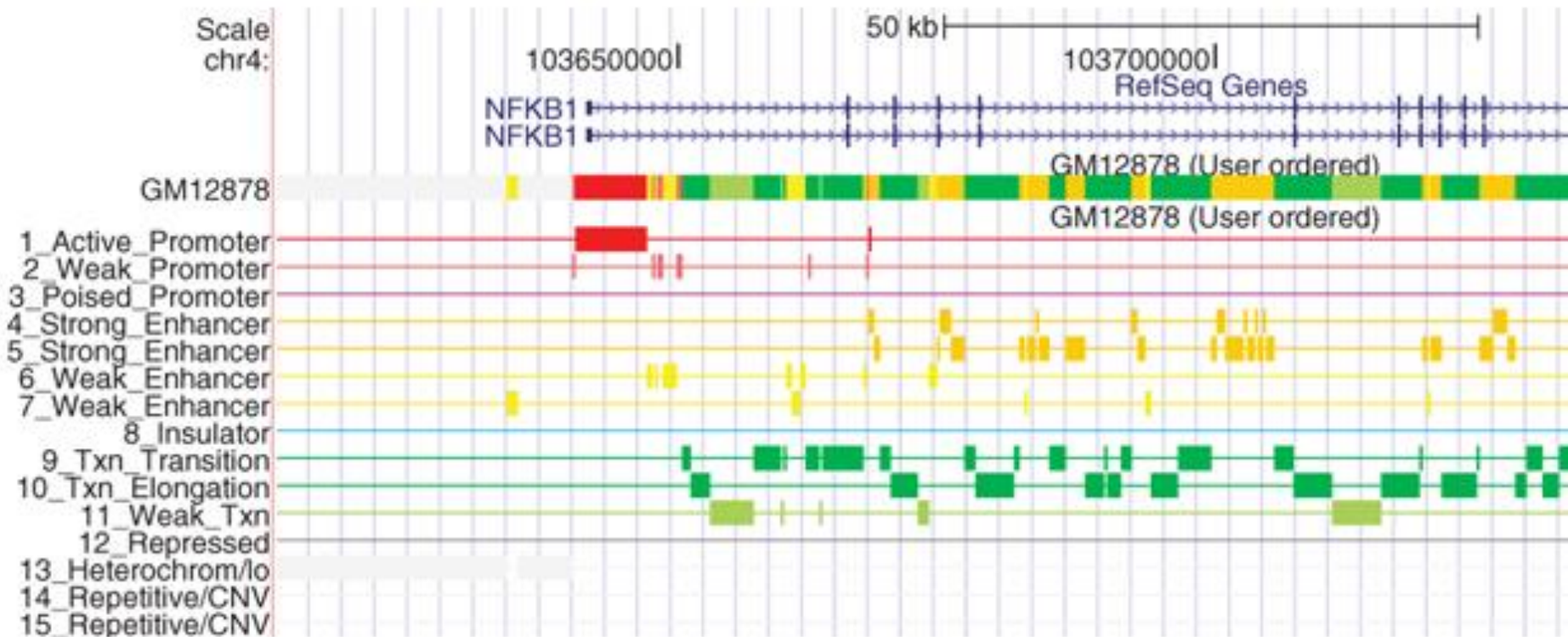
Large-scale epigenetic maps

- Epigenomes are condition-specific
- Roadmap Epigenomics Consortium and ENCODE surveyed over 100 types of cells and tissues



Genome annotation

- States are more interpretable than raw data

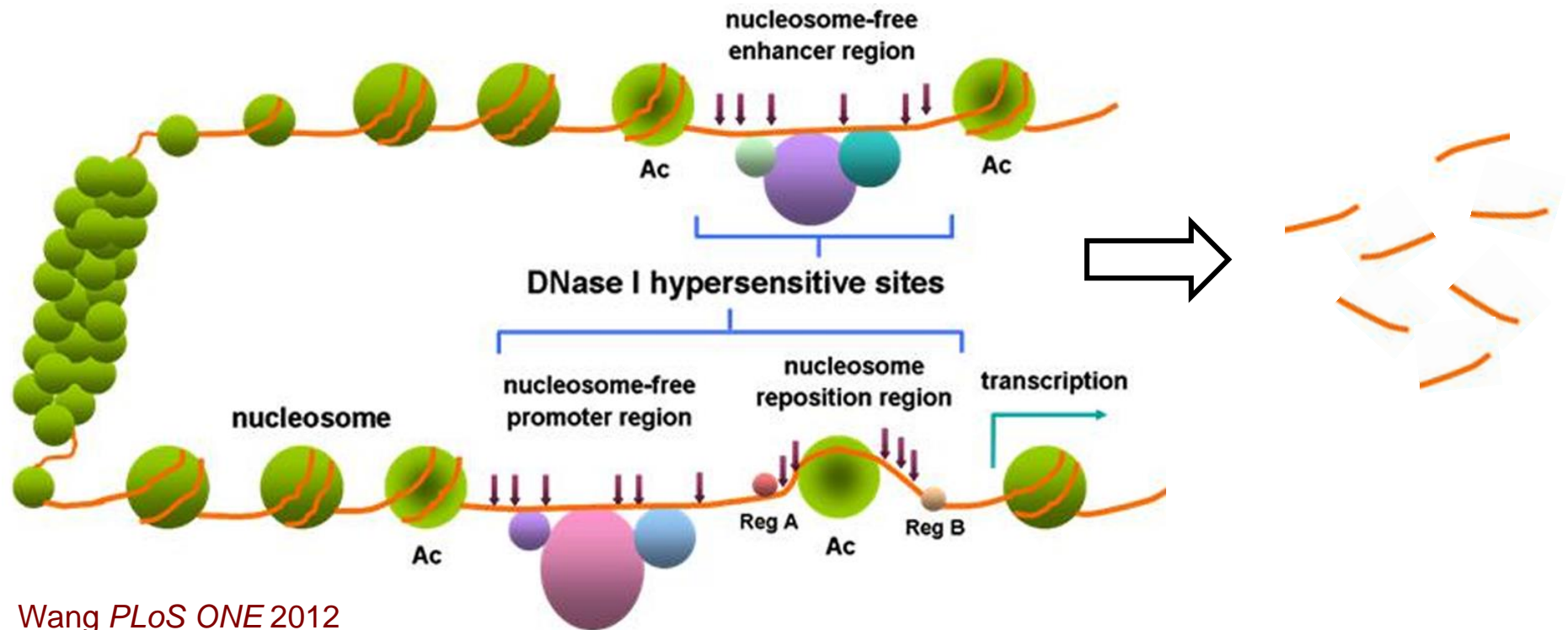


Ernst and Kellis *Nature Methods* 2012

Predicting TF binding with DNase-Seq

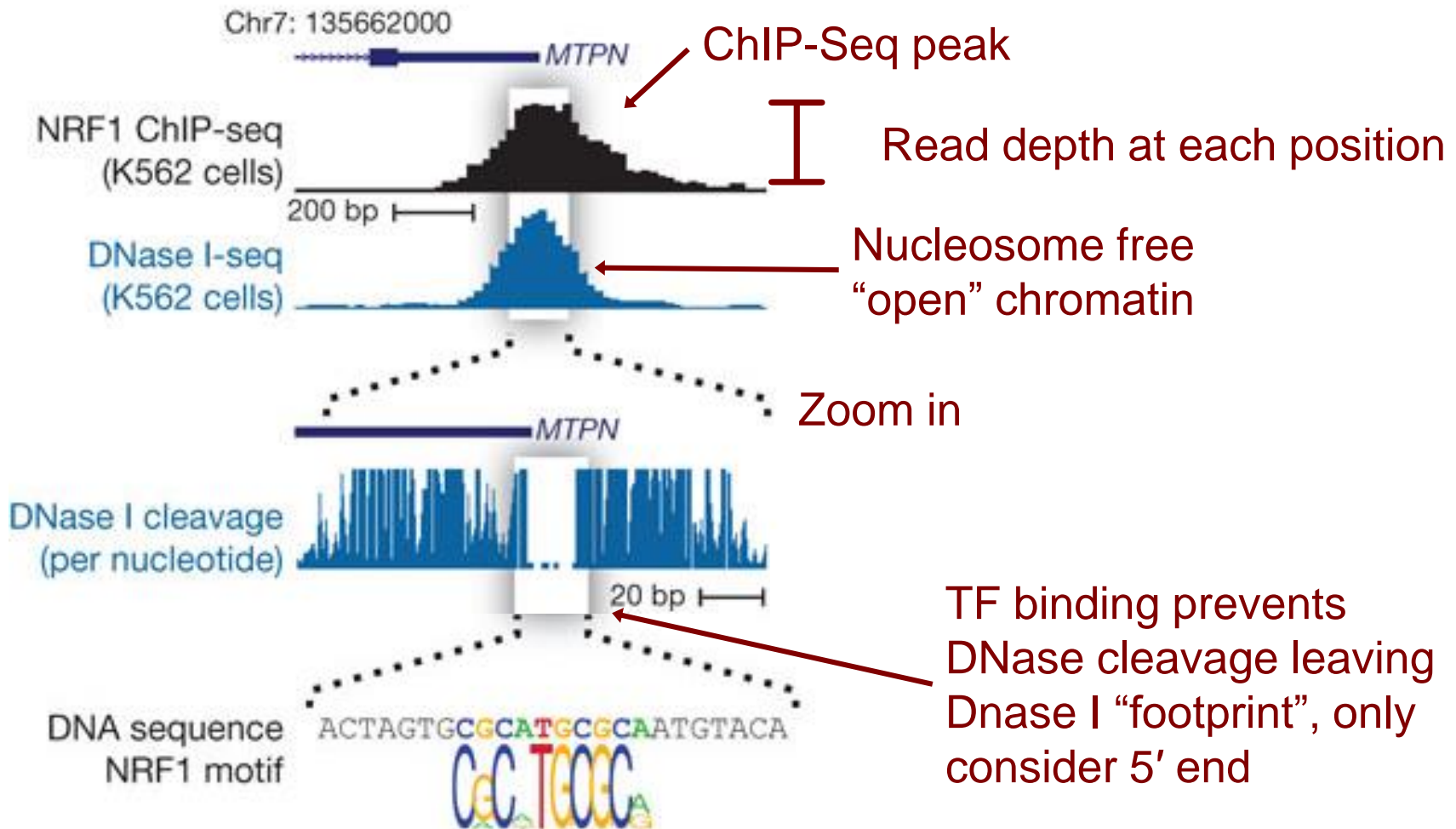
DNase I hypersensitive sites

- Arrows indicate DNase I cleavage sites
- Obtain short reads that we map to the genome



DNase I footprints

- Distribution of mapped reads is informative of open chromatin and specific TF binding sites

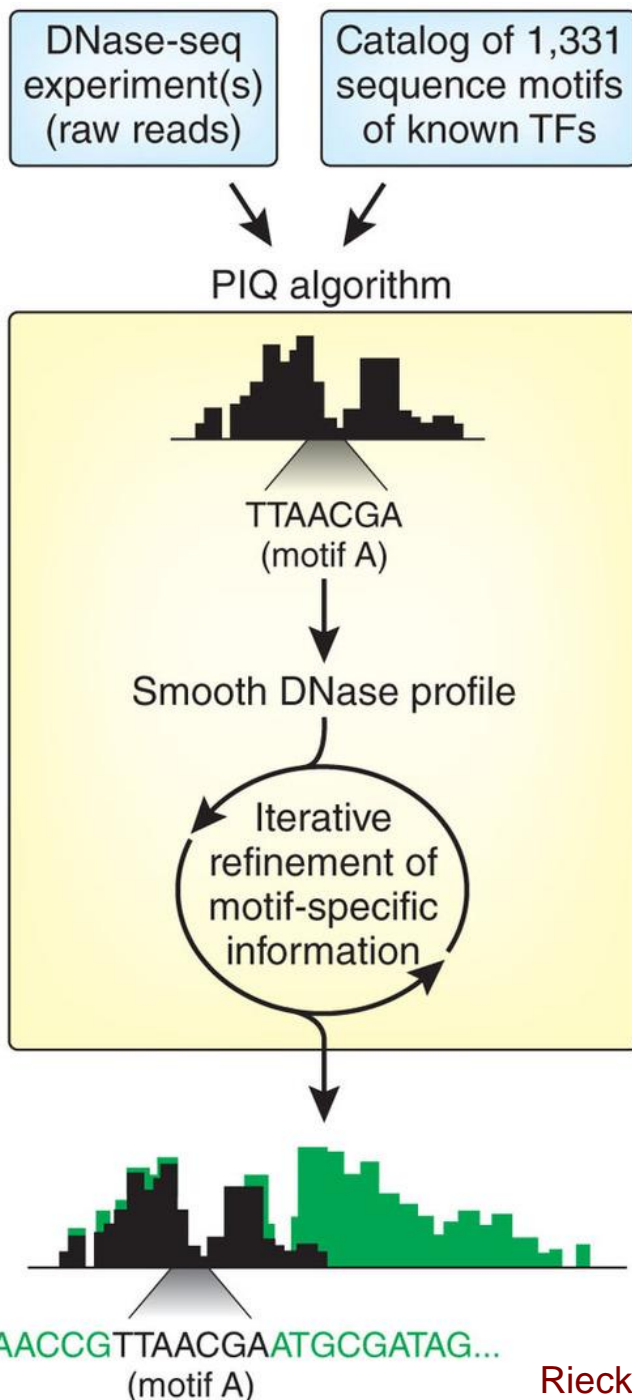


DNase I footprints to TF binding predictions

- DNase footprints suggest that **some** TF binds that location
- We want to know **which** TF binds that location
- Two ideas:
 - Search for DNase footprint patterns, then match TF motifs
 - Search for motif matches in genome, then model proximal DNase-Seq reads

← We'll consider this approach

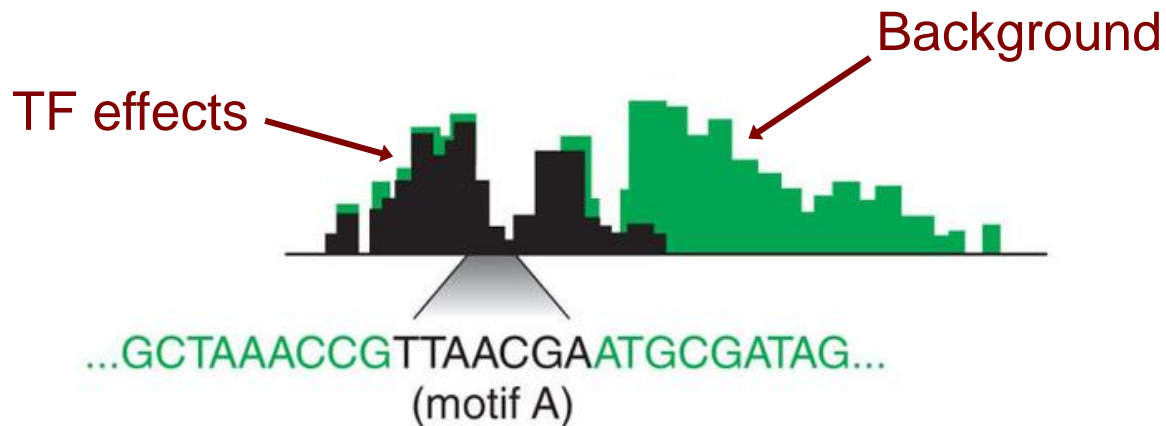
Protein Interaction Quantification (PIQ)



- Sherwood et al. *Nature Biotechnology* 2014
- **Given:** TF motifs and DNase-Seq reads
- **Do:** Predict binding sites of each TF

PIQ main idea

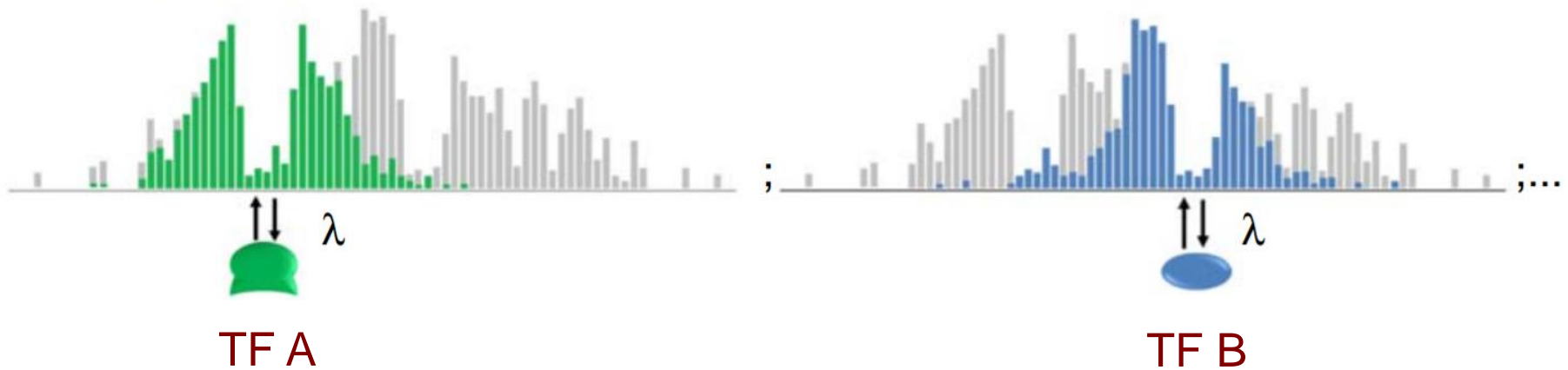
- With no TF binding, DNase-Seq reads come from some background distribution
- TF binding changes read density in a *TF-specific way*



PIQ main idea

- Shape of DNase peak and footprint depend on the TF

TF binding estimation



Sherwood *Nature Biotechnology* 2014

PIQ features

- We'll discuss
 - Modeling the DNase-Seq background distribution
 - How TF binding impacts that distribution
 - Priors on TF binding
- We'll skip
 - Modeling multiple replicates or conditions, cross-experiment and cross-strand effects
 - Expectation propagation
 - TF hierarchy: pioneers, settlers, migrants

Algorithm preview

- Identify candidate binding sites with PWMs
- Build a probabilistic model of the DNase-Seq reads
- Estimate TF binding effects
- Estimate which candidate binding sites are bound
- Predict pioneer, settler, and migrant TFs

DNase-Seq background

- Each replicate is noisy, don't want to over-interpret this noise
 - Only counting density of 5' ends of reads
- Manage two competing objectives
 - Smooth some of the noise
 - Don't destroy base pair resolution signal

Gaussian processes

- Can model and smooth sequential data
- Bayesian approach
- [Jupyter notebook demonstration](#)

TF DNase profile

- Adjust the log-read rate by a TF-specific effect at binding sites

$$\hat{\mu}_l = \mu_i + \begin{cases} \beta_{i-j,l} & |y_m - j| \leq W \text{ and } I_m = 1 \\ 0 & \text{otherwise} \end{cases}$$

DNase log-read rate adjusted for binding of factor l

DNase log-read rate at position i from Gaussian process

DNase profile for factor l

Location of binding site m

Window size

Whether site m is bound

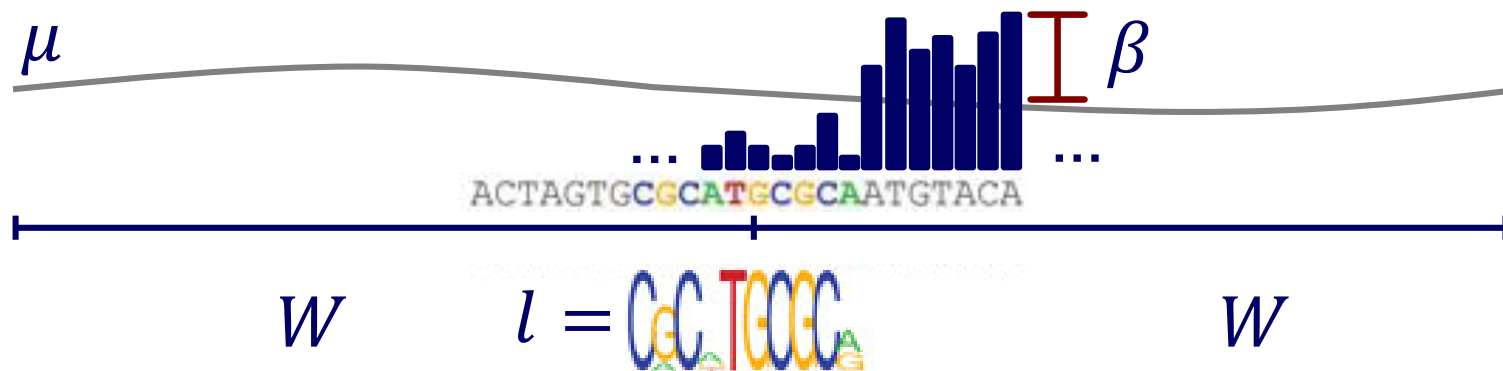
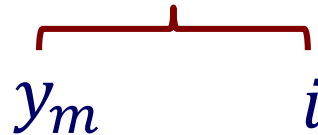
TF DNase profile

- DNase profiles represented as a vector for each TF

$$\hat{\mu}_l = \mu_i + \begin{cases} \beta_{i-j,l} & |y_m - j| \leq W \text{ and } I_m = 1 \\ 0 & \text{otherwise} \end{cases}$$

DNase profile for factor l

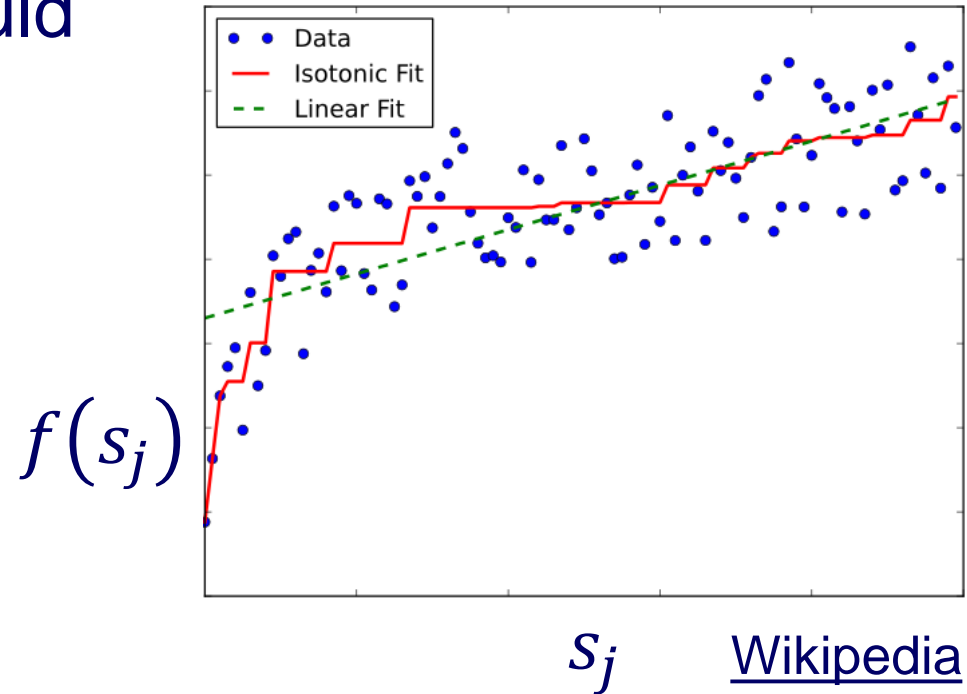
Can't be too far apart



Priors on TF binding

- TF binding event I_j should be more likely when
 - motif score s_j is high
 - DNase counts c_j are high
- Isotonic (monotonic) regression

Example only, not realistic data



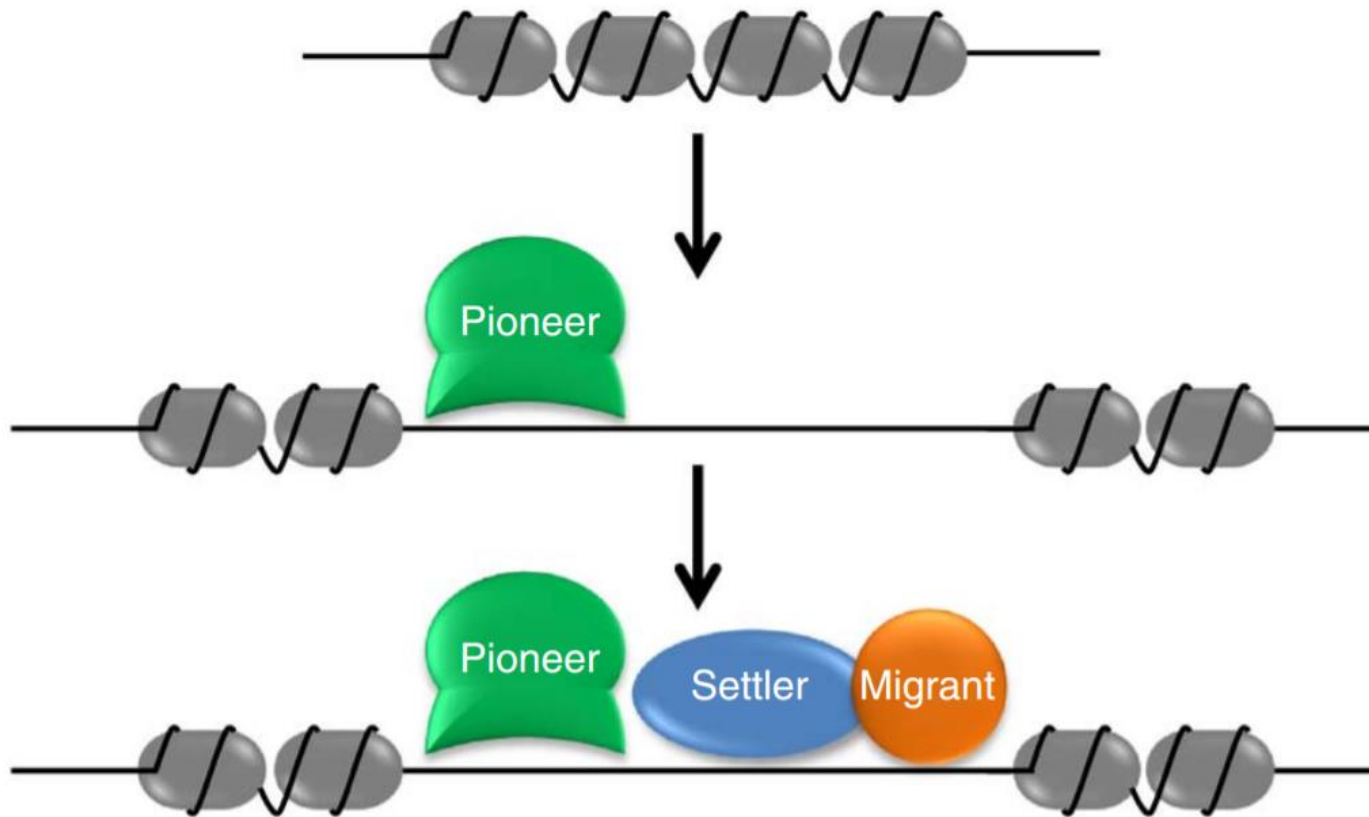
$$\log(P(I_j = 1)) = f(s_j) + g(c_j)$$

Full algorithm

- **Given:** TF motifs and DNase-Seq reads
- **Do:** Predict binding sites of each TF
- Identify candidate binding sites with PWMs
- Fit Gaussian process parameters for background
- Estimate TF binding effects $\beta_{i-j,l}$
- Iterate until parameters converge
 - Estimate Gaussian process posterior with expectation propagation
 - Estimate expectation of which candidate binding sites are bound
 - Update monotonic regression functions for binding priors

TF binding hierarchy

- Pioneer, settler, and migrant TFs



Sherwood *Nature Biotechnology* 2014

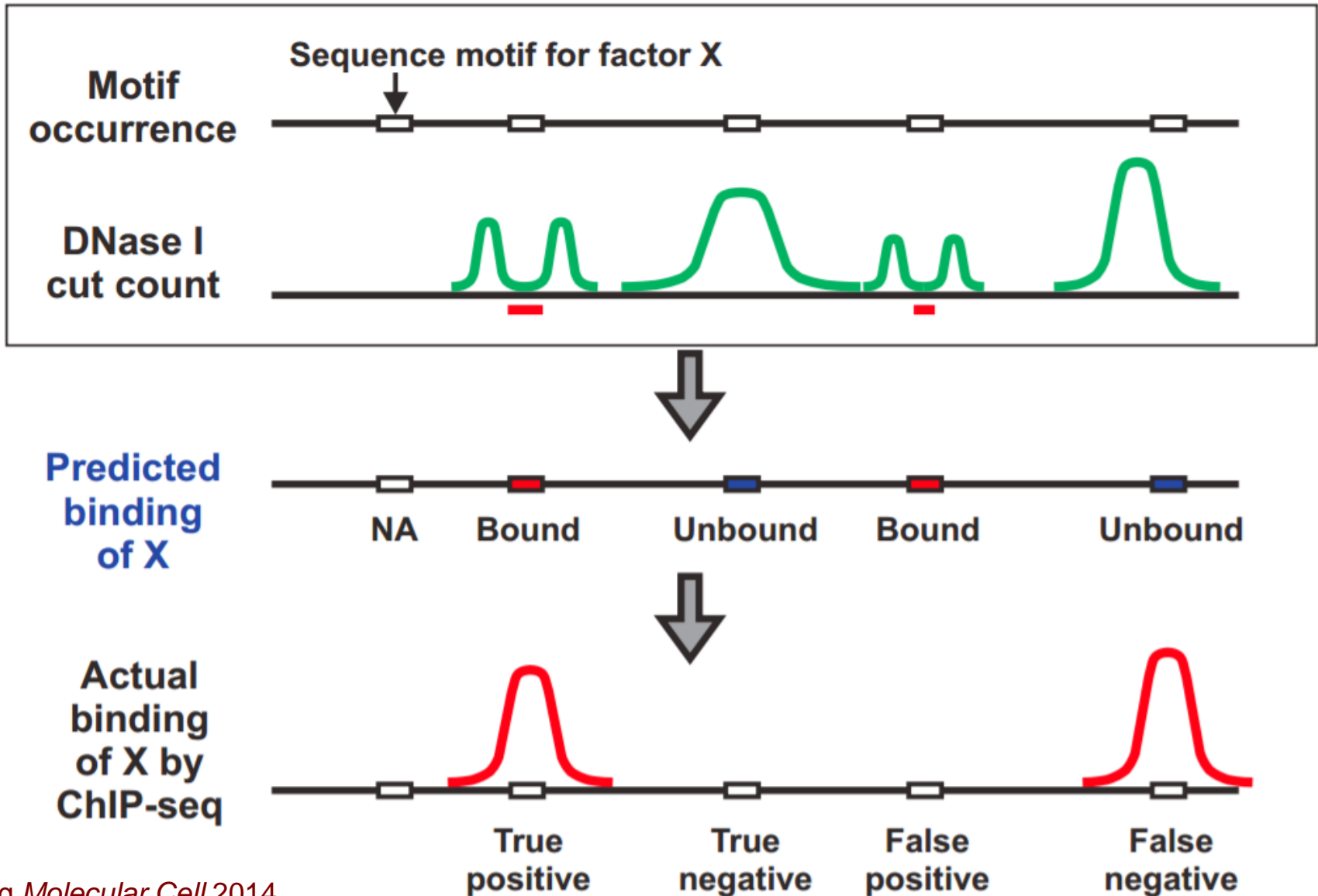
Evaluation: confusion matrix

- Compare predictions to actual ground truth (gold standard)

		Predicted	
		+	-
Actual	+ ●	TP	FN Type II error
	- ●	FP Type I error	TN

Lever *Nature Methods* 2016

Evaluation: ChIP-Seq gold standard



Evaluation: ROC curve

- Calculate receiver operating characteristic curve (ROC)
- True Positive Rate versus False Positive Rate
- Summarize with area under **ROC** curve (AUROC)

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

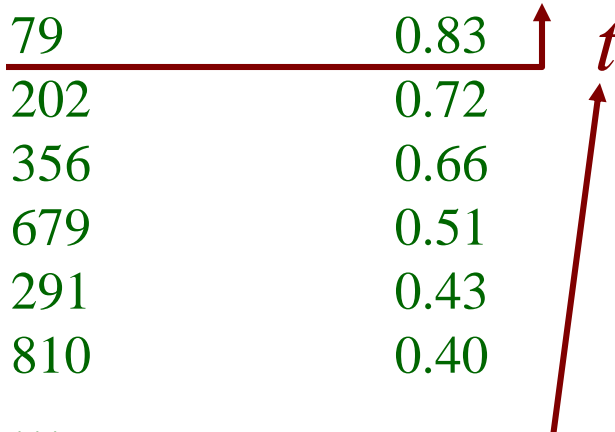
Includes true negatives

Reason to prefer precision-recall for class imbalanced data

Evaluation: ROC curve

- TPR and FPR are defined for a **set** of positive predictions
- Need to threshold continuous predictions
- Rank predictions
- ROC curve assesses all thresholds

Candidate binding site	$P(\text{bound})$	
764	0.99	Positive predictions
47	0.96	
942	0.91	
157	0.87	
79	0.83	
202	0.72	Negative predictions
356	0.66	
679	0.51	
291	0.43	
810	0.40	
...		

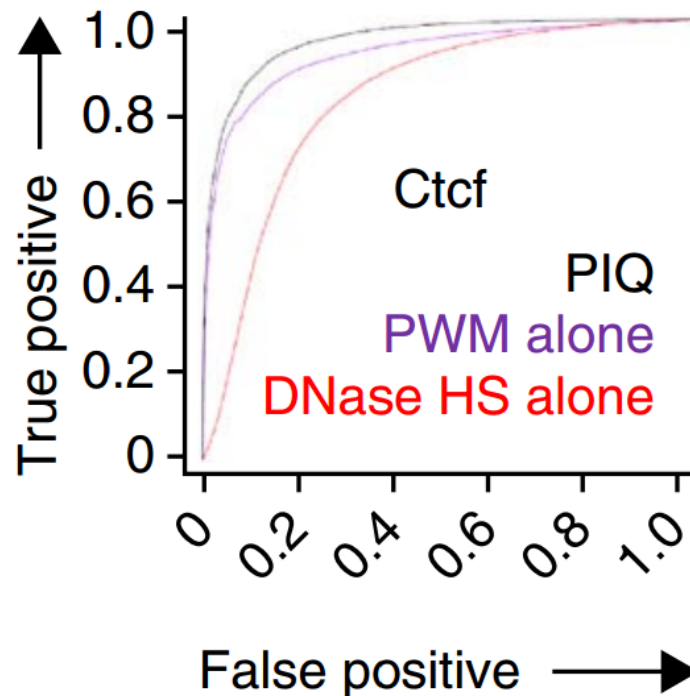


The diagram illustrates the thresholding process. A horizontal line is drawn at the probability value of 0.83. The data points above this line (764, 47, 942, 157, 79) are labeled as 'Positive predictions'. The data points below the line (202, 356, 679, 291, 810) are labeled as 'Negative predictions'. A vertical arrow labeled 't' points to the 0.83 threshold value on the probability axis.

Calculate TPR and FPR at all thresholds t

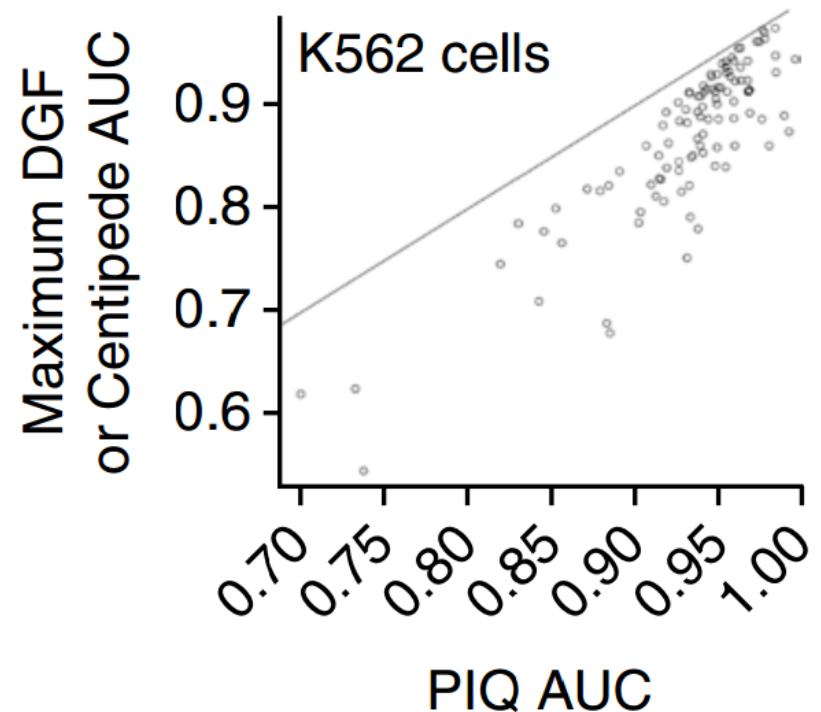
PIQ ROC curve for mouse Ctcf

- Compare predictions to ChIP-Seq
- Full PIQ model improves upon motifs or DNase alone



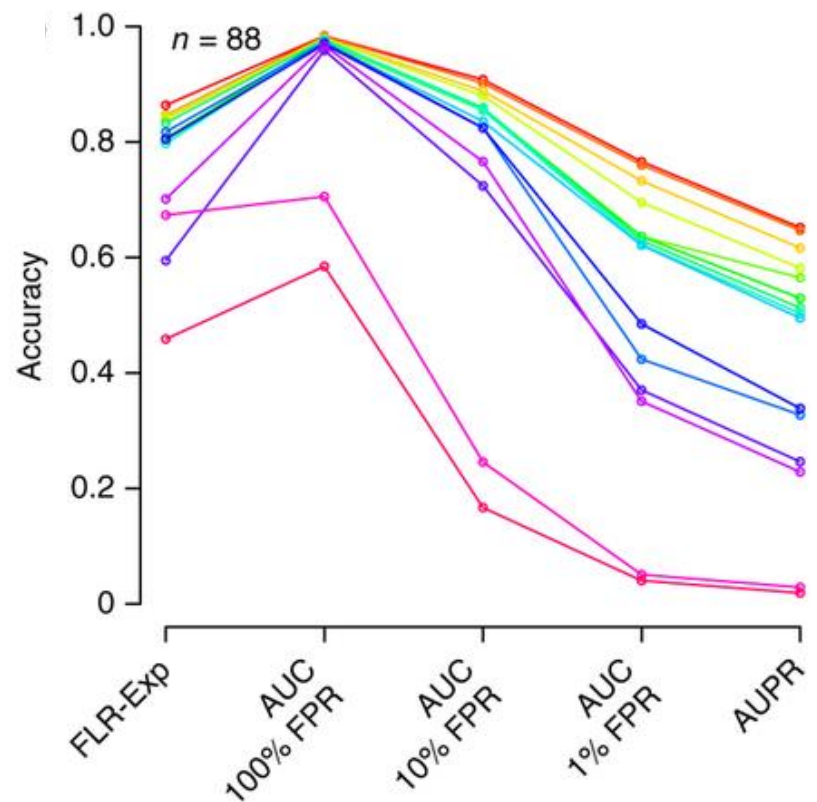
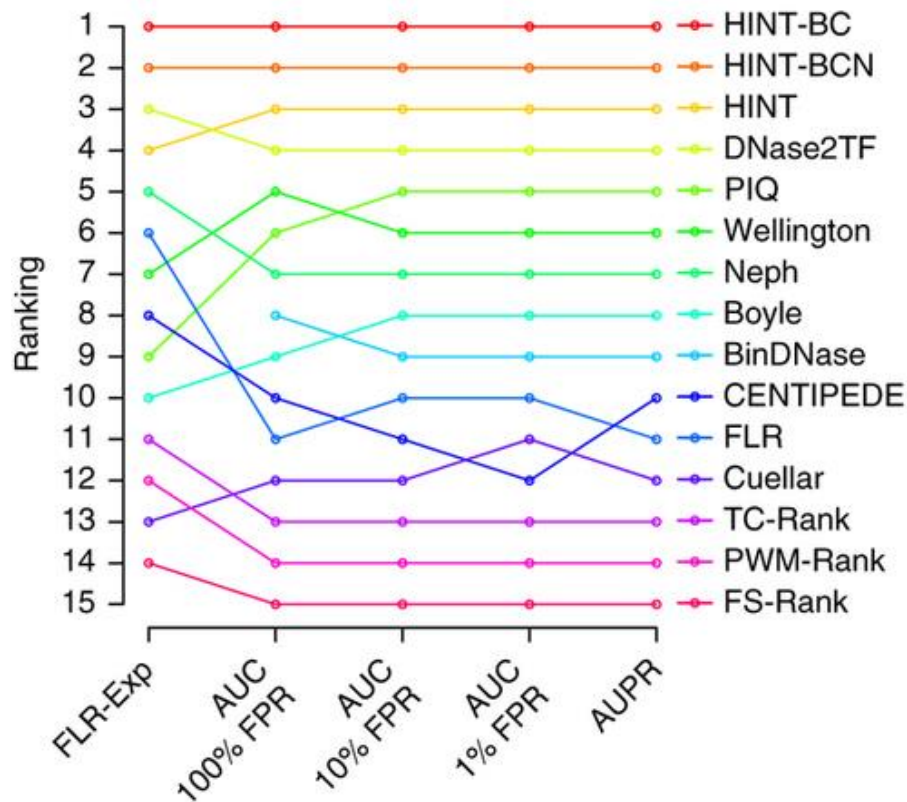
PIQ evaluation

- Compare to two standard methods
 - 303 ChIP-Seq experiments in K562 cells
 - Centipede, digital genomic footprinting
- Compare AUROC
 - PIQ has very high AUROC
 - Mean 0.93
 - Corresponds to recovering median of 50% of binding sites

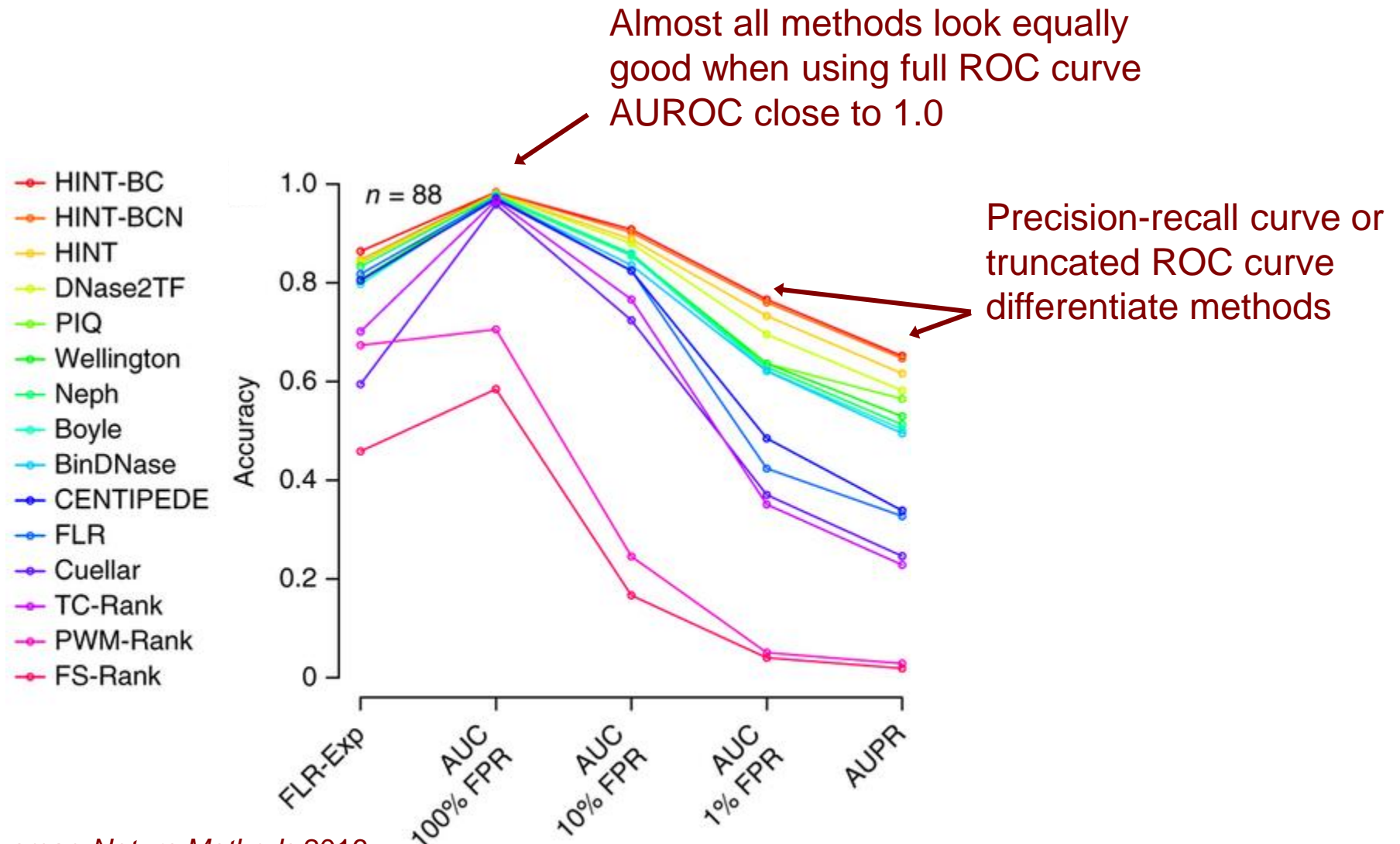


DNase-Seq benchmarking

- PIQ among top methods in large scale DNase benchmarking study
- HMM-based model HINT was top performer



Downside of AUROC for genome-wide evaluations



PIQ summary

- Smooth noisy DNase-Seq data without imposing too much structure
- Combine DNase-Seq and motifs to predict condition-specific binding sites
- Supports replicates and multiple related conditions (e.g. time series)